ehp ENVIRONMENTAL HEALTH PERSPECTIVES

ehponline.org

Performance in Omics Analyses of Blood Samples in LongTerm Storage: Opportunities for the Exploitation of Existing Biobanks in Environmental Health Research

Dennie G.A.J. Hebels, Panagiotis Georgiadis, Hector C. Keun, Toby J. Athersuch, Paolo Vineis, Roel Vermeulen, Lützen Portengen, Ingvar A. Bergdahl, Göran Hallmans, Domenico Palli, Benedetta Bendinelli, Vittorio Krogh, Rosario Tumino, Carlotta Sacerdote, Salvatore Panico, Jos C.S. Kleinjans, Theo M.C.M. de Kok, Martyn T. Smith and Soterios A. Kyrtopoulos

http://dx.doi.org/10.1289/ehp.1205657

Online 05 February 2013

NIEHS
National Institute of
Environmental Health Sciences

National Institutes of Health
U.S. Department of Health and Human Services

Performance in Omics Analyses of Blood Samples in Long-Term Storage: Opportunities for the Exploitation of Existing Biobanks in Environmental Health Research

Authors: Dennie G.A.J. Hebels[1*], Panagiotis Georgiadis[2*], Hector C. Keun[3,4], Toby J. Athersuch[3,4], Paolo Vineis[3],Roel Vermeulen[5], Lützen Portengen[5], Ingvar A. Bergdahl[6], Göran Hallmans[7], Domenico Palli[8], Benedetta Bendinelli[8], Vittorio Krogh[9], Rosario Tumino[10], Carlotta Sacerdote[11,12], Salvatore Panico[13], Jos C.S. Kleinjans[1], Theo M.C.M. de Kok[1], Martyn T. Smith[14] and Soterios A. Kyrtopoulos[2], on behalf of the EnviroGenomarkers project consortium

Affiliations: [1]Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands; [2]National Hellenic Research Foundation, Institute of Biology, Medicinal Chemistry and Biotechnology, Athens, Greece; [3]Imperial College London, MRC-HPA Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, London, UK; [4] Imperial College London, Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, London, UK; [5]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands; [6]Occupational and Environmental Medicine, Department of Public Health and Clinical Medicine and Department of Biobank Research, Umeå University, Umeå, Sweden; [7]Nutrition Research, Department of Public Health and Clinical Medicine and Department of Biobank Research, Umeå University, Umeå, Sweden; [8]The Institute for Cancer Research and Prevention, Italy; [9]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; [10]Cancer Registry and Histopathology Unit, "Civile - M.P. Arezzo" Hospital, Ragusa, Italy; [11]Center for Cancer Prevention (CPO-Piemonte), Turin, Italy;

[12]Unit of Epidemiology and Molecular Genetics, Human Genetics Foundation (HuGeF), Turin, Italy; [13]Department of Clinical and Experimental Medicine, Federico II University, Naples, Italy. **[14]**Genes and Environment Laboratory, School of Public Health, University of California, Berkeley, California, USA

* contributed equally to this work

Corresponding author:

Soterios A. Kyrtopoulos, National Hellenic Research Foundation, Institute of Biology, Medicinal Chemistry and Biotechnology, 48 Vas. Constantinou Ave., Athens 11635, Greece

tel: +30-210-7273740; fax: +30-210-7273677; email: skyrt@eie.gr

Additional members of the EnviroGenomarkers consortium:

Ralph Gottschalk[1], Danitsja van Leeuwen[1], Leen Timmermans[1], Maria Botsivali[2], Christina Papadopoulou[2], Aristotelis Chatziioannou[2], Ioannis Valavanis[2], Paolo Vineis[3], Marc Chadeau-Hyam[3], Rachel Kelly[3], Fatemeh Saberi-Hosnijeh[5], Beatrice Melin[15], Per Lenner[15], Manolis Kogevinas[16], Euripides G. Stephanou[17], Antonis Myridakis[17], Lucia Fazzo[18], Marco De Santis[18], Pietro Comba[18], Hannu Kiviranta[19], Panu Rantakokko[19], Riikka Airaksinen[19], Päivi Ruokojärvi[19], Mark Gilthorpe[20], Sarah Fleming[20], Thomas Fleming[20], Yu-Kang Tu[20], Bo Jonsson[21], Thomas Lundh[21], Wei J. Chen[22], Wen-Chung Lee[22], Chuhsing Kate Hsiao[22], Kuo-Liong Chien[22], Po-Hsiu Kuo[22], Hung Hung[22], Shu-Fen Liao[22]

Affiliations: [15]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden; [16]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; [17]Environmental Chemical Processes Laboratory, University of Crete, Heraklion, Greece;

[18]Istituto Superiore di Sanita, Rome, Italy; [19]National Institute for Health and Welfare, Kuopio, Finland; [20]University of Leeds, UK; [21]Lund University, Lund, Sweden; [22]National Taiwan University, Taipei, Taiwan

Running title: omics analysis of biosamples in long-term storage

Keywords: biomarkers, epigenomics, metabolomics, metabonomics, molecular epidemiology, proteomics, transcriptomics

Acknowledgements: Research support by the European Union (Grant number 226756). Epigenomics analyses were conducted under contract by CBM (Cluster in Biomedicine) S.c.r.l., Trieste, Italy, an Illumina Certified Service Provider. The authors wish to thank for their technical contributions Margarita Bekyrou, Stella Kaila and Nena Burger.

Disclaimer of competing interests: The authors declare they have no competing financial interests.

Abbreviations: ANOVA, Analysis of Variance; FDR, False Discovery Rate; NSHDS, North Sweden Health and Disease Study; PCA, Principal Component Analysis; RIN, RNA Integrity Number; RSD, Relative Standard Deviation; STEM, Short Time-series Expression Miner; QC, Quality Controls; UPLC-ToFMS, Ultra Performance Liquid Chromatography - Time-of-flight Mass Spectrometry

Abstract

Background: The suitability for omic analysis of biosamples collected in previous decades and currently stored in biobanks is not known.

Objectives: We evaluated the influence of handling and storage conditions of blood-derived biosamples on transcriptomic, epigenomic (CpG methylation) and plasma metabolomic (UPLC-ToFMS) and wide-target proteomic profiles.

Methods: Initially we collected fresh blood samples without RNA preservative in heparin, EDTA or citrate and held them at room temperature for up to 24hr prior to fractionation into buffy coat, erythrocytes and plasma and freezing at -80^{o}C or in liquid nitrogen. We developed methodology for RNA isolation from the buffy coats and conducted omic analyses. Finally, we analysed analogous samples from the EPIC-Italy and Northern Sweden Health and Disease Study biobanks.

Results: Microarray-quality RNA could be isolated from buffy coats (including most biobank samples) frozen within 8hr of blood collection, by thawing in RNA preservative. Different anticoagulants influenced the metabolomic, proteomic and, to a lesser extent, transcriptomic profiles. The latter were most affected by the delay (as little as 2hr) prior to blood fractionation, while storage temperature had minimal impact. Effects on metabolomic and proteomic profiles were noted in samples processed 8hr or more after collection, but none due to storage temperature. None of the variables examined significantly influenced the epigenomic profiles. No systematic influence of time-in-storage was observed in samples stored over a period of 13-17 years.

Conclusions: Most samples currently stored in biobanks are amenable to meaningful omics analysis, provided that they satisfy collection and storage criteria defined in this study.

Introduction

The use of omics technologies has led to improved understanding of the mechanisms of toxicity and to new knowledge of value for environmental health research (Ellinger-Ziegelbauer 2009; McHale et al. 2010). By providing global and quantitative information on changes in critical cellular components under the influence of environmental factors, omics profiling greatly facilitates the discovery of biomarkers and is seen as a key tool in the development of the concept of the exposome (Rappaport and Smith 2010).

The application of omics technologies in epidemiological studies raises certain practical issues of sample suitability, especially in relation to RNA quality for transcriptomics analysis, requiring that care be taken for blood samples to be collected and stored in the presence of RNA preservative. However, millions of human biosamples currently in cold storage in older biobanks were collected and processed by methods that did not anticipate the demands of omics technologies. Such biobanks represent a precious resource for environmental health research, especially in view of the fact that newly constructed biobanks will take many years to accrue enough cases of chronic diseases in their prospective cohorts to allow relevant biomarker research. Yet no study has evaluated systematically the influence on omic profiles of the handling and prolonged storage of blood samples and their components in these biobanks.

In the context of the European project EnviroGenomarkers (EnviroGenomarkers project, n.d.) blood-derived biobank samples are being analysed on multiple omic platforms with the aim of discovering new biomarkers of exposure and disease risk. As a first step in this project a study was conducted, whose results are reported here, to evaluate the reliability of omics data obtained from archived biosamples collected prior to the advent of omics technologies.

Methods

The omics technologies employed include transcriptomics, epigenomics (CpG methylation) and plasma UPLC-ToFMS metabolomics. In addition, a multi-analyte profiling platform was used as a tool for a wide-target plasma proteomics screen. All international regulations regarding the use of human subjects were complied with. Ethical approval for the use of volunteers was obtained from the research ethic committees of the University of Maastricht and the National Hellenic Research Foundation and written informed consent was obtained from all volunteers prior to the study. The use of biobank samples was approved by the corresponding ethical committees.

During Phase I of the study, we established methods for the isolation of RNA of the desired quality from buffy coats isolated from blood freshly collected and processed without RNA preservative. We also evaluated the influence on omics profiles of sample handling and storage-related parameters selected following scrutiny of the procedures employed at the biobanks participating in study. The results obtained were used to establish minimum criteria that samples must satisfy in order to be suitable for reliable omics analysis. During Phase II, we analysed historic samples satisfying these criteria, stored in the EPIC-Italy and the North Sweden Health and Disease Study (NSHDS) biobanks (Bingham and Riboli 2004; Hallmans et al. 2003) so as to evaluate the influence of long-term storage.

Sample collection

Phase I: We collected fresh blood from healthy volunteers using different anticoagulants (heparin, EDTA or citrate) and processed it in different ways. For practical reasons we conducted several blood collection experiments, in the context of which different variables were evaluated (for details see Supplemental Material, Methods, Design of Phase I experiments). After allowing the blood samples to stand at room temperature for different times up to 24hr ("bench-time"), we separated buffy coats, erythrocytes and plasma by centrifugation for 15 minutes at 1,500g at room temperature, followed by aliquoting and immediate storage at -80^{o}C or in liquid nitrogen. To control for effects of interindividual variation, in 1 experiment we collected blood from 1 person in each of the 3 anticoagulants, processed it for fractionation and stored it in both liquid nitrogen and at -80^{o}C but without variation in bench-time.

The duration of cold storage of the blood fractions prior to omics analysis varied from several weeks to several months. We conducted full-scale metabolomics and wide-target proteomics analysis on all samples from a single blood collection experiment, in the context of which we evaluated all combinations of the parameters of interest (donors, bench-times, anticoagulants, storage temperature). On the other hand, for practical reasons we generally conducted transcriptomics and epigenomics analyses aimed at evaluating the influence of individual variables on a more limited number of samples.

Phase II: We used biosamples from the participating biobanks, satisfying the cut-off criteria established during Phase I, to evaluate the quality of extracted RNA and DNA and carry out omics analysis. Samples from EPIC-Italy used citrate as anticoagulant and had been stored in cryostraws in liquid nitrogen for 11-19 years. Their recorded collection-to-storage times were 55-347 min. Samples from NSHDS contained heparin or EDTA as anticoagulant and had been stored in plastic cryovials at -80^{o}C for 4-19 years. Their collection-to-storage time was always shorter than 1hr. To evaluate the impact of storage time on the different omics profiles, we analysed samples from the same set of 31 subjects from each biobank. To minimise the effect of variables other than storage time, these samples were selected to come only from healthy female donors and from the same collection centre per biobank. Their storage time prior to analysis was 13-17 years, while the collection-to-storage times for the EPIC-Italy sub-set were 100-198 min.

RNA and DNA isolation

To establish methods for RNA extraction from buffy coats stored in the absence of RNA preservative, we thawed Phase I samples fully immersed in RNAlater or Qiazol (Qiagen, Venlo, The Netherlands) and subsequently extracted RNA according to the manufacturer's instructions. We quantified RNA with a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and used an Agilent 2100 Bioanalyzer (Agilent Technologies, Amstelveen, The Netherlands) to assess its quality, including RIN which represents the degree of RNA fragmentation (Schroeder at al., 2006). All RIN (RNA integrity number) values were >6, as required for good-quality microarray-based analysis. While the above procedures also allow extraction of microRNA, this was not systematically assessed in these samples.

In Phase II, we adapted the RNA extraction methodology developed in Phase I for use with biobank samples. For this, we handled all samples individually and immediately after retrieval from storage. We divided sample-containing straws from EPIC-Italy for different applications by cutting with RNase-free tools on a stainless steel plate imbedded in a box of dry ice to prevent thawing during handling, and pushed out half of the frozen buffy coat with a thin stainless steel plunger directly into 1.2 ml of the RNAlater (Qiagen) solution. The other half was used for DNA isolation. We retrieved samples from NSHDS from their cryovials in a frozen state by making a small opening at the bottom of the vial using a hot plunger and pushing the sample out with another plunger. After subdividing the buffy coat on a dry ice-cooled steel plate using a RNase-free scalpel, we immediately thawed the part destined for RNA extraction in 1.2 ml of RNAlater (Qiagen) (see Supplemental Material for video of procedures). We replaced the remaining pellet in a cryovial and immediately returned it to cold storage for later isolation of DNA. RNA was isolated on the same day using the RiboPure^{TM} Blood kit (Ambion, Austin, TX, USA) with the miRNA isolation protocol.

We used buffy coats free of RNA preservative for DNA isolation since material thawed in the presence of RNA-later or Qiazol (Qiagen) proved impossible to dissolve for DNA isolation. We thawed the samples on ice and isolated DNA using the QIAamp Blood Mini Kit (Qiagen), evaluating it spectrophotometrically and by agarose gel electrophoresis.

Transcriptomics

We conducted Agilent 4x44K human whole genome microarray analyses by standard methodology. Briefly, we reverse transcribed each RNA sample into cDNA and labelled it with cyanine 3 prior to hybridization. Subsequently we washed the slides and scanned them an Agilent Technologies G2565CA DNA Microarray Scanner.We established the technical performance and quality of the microarrays by visual evaluation of the scan images before and after within- and between-array normalization (LOESS and A-quantile, respectively). We imputed missing values in GenePattern (version 3.1) using the *k* nearest neighbors approach (k = 15, Euclidian metric). For more details on the transcriptomics and other omics methodologies employed see Supplemental Material, Methods.

Epigenomics

We conducted genome-wide analysis of DNA methylation using Infinium HumanMethylation450 BeadChips (Illumina, San Diego, CA, USA) which contain 485,764 probes (>99% with CpG dinucleotides), following the experimental protocol recommended by the manufacturer. We preprocessed the data with the GenomeStudio (2011.1) Methylation module (1.9) (Illumina) and evaluated them using an adaptation of HumMeth27QCReport (Mancuso et al. 2011). We used Gene ARMADA (Chatziioannou et al. 2009) for within- and between-array normalization (linear LOESS and A-quantile, respectively) and imputation of missing values (*k* nearest neighbors approach).

Metabolomics

We analysed plasma samples by UPLC-ToFMS after deproteinization with methanol. We conducted reverse-phase chromatography on an Acquity UPLC system (Waters Corporation, Milford, MA, USA) with a C_{18} column (Waters) and binary gradient elution (20%-100% acetonitrile/water in ~25mins). Online analysis of the eluent was performed using a quadrupole time-of-flight mass spectrometer (Waters), with data collected in centroid mode in the m/z range 100-1000. In Phase II, we prepared samples in batches by biobank. Data were processed using Databridge & XCMS.

Wide-target plasma proteomics

We conducted targeted proteomic analysis of plasma samples using the Lab-MAP multi-analyte profiling technology (Luminex, Austin TX, USA). We analysed Phase I samples for interleukin (IL)2, IL6, IL8, IL10, and tumor necrosis factor alpha as previously described (Saberi-Hosnijeh et al. 2010), while we analysed Phase II samples for an additional 23 proteins related to immune responses (for a full list see Supplemental Material, Methods, Wide-target plasma protomics) according to the manufacturer's protocol. Phase I and II samples were run in a single batch on a single plate. Non-detectable concentrations (<1.22 pg/ml for all analytes) were imputed via a maximum likelihood estimation method (Lubin et al. 2004).

Statistical evaluation

The data were evaluated using principal component analysis (PCA), ANOVA, paired *t*-test, mixed effect models, relative standard deviation (RSD = standard deviation/mean), false discovery rate (FDR, Storey's q-value) and STEM (Short Time-series Expression Miner; Ernst and Bar-Joseph 2006) analysis. PCA plots were used to visualise the impact of different sample handling parameters on omics signals as reflected in the variation of the different principal components (PC's). STEM analysis allows the identification of significant temporal trends in expression profiles and the genes associated with them. Because of the severe heteroscedasticity of β-values at highly methylated or unmethylated CpG sites, M values [$M=\log_2$(methylated/unmethylated)] were used for statistical analysis of DNA methylation data (Du et al. 2010).

Results

Transcriptomics

Phase I: RNA quality and quantity were both significantly ($p<1x10^{-5}$) higher in buffy coat samples thawed in the presence of RNAlater as compared to Qiazol (RIN: 7.17±0.51 vs. 6.14±0.72; RNA yield: 6.03±2.16 vs. 2.25±1.04 μg), and for this reason the former was employed routinely. No systematic effect of bench-time, anticoagulant or storage temperature on RIN values was observed (Table 1). RNA yield was unaffected by bench-time and was higher for citrate samples regardless of storage temperature ($p<0.01$, possibly due to minor interference of heparin and EDTA in the RNA extraction procedure) and for -80°C samples regardless of anticoagulant ($p<0.05$). We confirmed these findings using blood samples originating from one single individual with different anticoagulants and a bench-time of 0hr (results not shown).

We performed transcriptomics analysis of the effects of donor and bench-time on material from 4 subjects. Genes with more than one flagged/missing time-point for any subject were completely filtered out of the dataset, leaving 27,181 genes. Plots of principal components (PC) according to the various sample-related parameters (Figure 1A,B) shows clear separation between the subjects (except for one time-point of one subject), based on PC 1-3, while a bench-time-dependent trend is observed up to 8hr in PC4 (a bench-time of 24hr was omitted since a small-scale RT-PCR experiment already showed substantial gene expression changes at this time point - results not shown). We investigated this trend further by performing an ANOVA across the 4 time points and using the resulting 3,372 significant genes ($p <0.05$) in a STEM analysis to identify significantly represented temporal gene expression profiles. Two significant profiles were identified, corresponding to a gradual decrease or increase in expression and comprising together 83% of the genes with significant differences in expression based on ANOVA, with a between-subject of overlap of 90% (Supplemental Material, Figure S1A/B). Time-point comparisons showed considerable numbers of differentially expressed genes (1,000-3,000) at all time points, their numbers roughly doubling in going from 2h to 4h (Supplemental Material, Figure S1C). A pathway analysis on the two significant temporal STEM profiles revealed mainly involvement of the biological processes apoptosis, stress signaling and DNA damage repair (results not shown).

A list of genes with significant differences in expression based on ANOVA (Bonferroni-corrected p<0.05) which may be suitable as bench-time effect markers is presented in Table S1.

For the anticoagulant and storage temperature analysis, again on groups of 4 subjects, all genes flagged in any subject were filtered out, leaving 28,478 and 27,552 genes, respectively. PCA shows again a clear separation between subjects, but also some distance between the three anticoagulants, especially heparin (Figure 1C). Paired *t*-test analysis shows significant differences between all 3 anticoagulants (Supplemental Material, Figure S1D), with the largest differences (though not as large as with bench-time) being found between heparin and either EDTA or citrate, both with and without a log2 ratio cut-off of 0.5. Differences in the gene expression pattern were also identified between samples stored at -80^{o}C and liquid nitrogen (Figure 1D), with 2,193 differentially expressed genes (551 genes with an additional 0.5log2 ratio cut-off), but the FDR stays relatively high (35%).

In order to compare the impact of sample processing-related variables to that due to assay technical variability, we used technical repeats (2-3 per subject) to determine the coefficient of variation of corresponding log2-expression signals (average of 2.7%). Same-individual bench-time variation for all but one time point comparison (4 vs. 8hr) and for EDTA vs. heparin was significantly higher than technical variation (ranging up to 4.2%), while for the other anti-coagulant and storage temperature comparisons the variation was not significantly different. This means that bench-time is the main source of sample processing-related variability, while the effects of the other two variables may be overshadowed by technical noise.

Phase II: Using the procedures described, adequate amounts of RNA with RIN>6.0 (average RIN=7. 2, similar to fresh Phase I samples) could be isolated from approximately 85% of the biobank samples extracted (64 from EPIC-Italy and 50 from NSHDS) (Table 2), with no observable systematic effect of storage time (results not shown). .

To test the performance of biobank samples in transcriptomics analysis, we initially used 4 EPIC-Italy samples to compare the technical quality of the microarray data with those obtained with 4 Phase I samples stored at -80°C (different donors, 2 heparin and 1 EDTA with bench-time 0hr, 1 heparin with bench-time 24hr). All RNAs were hybridized against freshly isolated RNA from Phase I samples. No differences could be seen between the quality of the arrays hybridized with fresh or biobank samples. After normalisation, a boxplot showed similar data distribution between all samples (equal medians) (Figure 1E). After filtering flagged features, the number of remaining high-quality probes showed no significant difference across the arrays (Figure 1F).

PCA of the transcriptomic profiles of 31 samples from each biobank, selected as described in Methods, does not suggest any consistent effect of a storage time within the range 13-17 years (Figure 1G,H). ANOVA across these samples showed only 14 and 76 genes for EPIC-Italy and NSHDS (out of a total of 29,662), respectively, to vary significantly ($p<0.0033$) according to storage time; however, the FDR level was around 100% meaning that these were most likely false positives. We could not make a meaningful evaluation of the effect of collection-to-storage time on the transcriptomic (or any other) profile because of the small range of variation of this variable among the samples analysed (100-198 min).

A comparison of 6 low-RIN samples (5.9 - 6.9) with 6 high-RIN samples (8.5-8.8) yielded only one differentially expressed gene at an FDR of 10% (results not shown), indicating that RNA quality was not a significant factor influencing the transcriptomic profiles of biobank samples. As an additional test of data quality, we evaluated the expression of 3 blood reference genes

(B2M, GAPDH, PPP1CA) and 11 immunomodulatory marker genes (CXCL1, HMOX1, ICAM1, IL1B, IL1RN, IL6R, MMP9, PTGS2, SERPINE1, TGFB1, TNF) (Karlovich et al 2009) in these and in Phase I samples (all bench-times, 4 subjects) (Table S2). All genes were expressed in all sample sets, with the log2-transformed intensities of the 3 reference genes and the majority of the immunomodulatory genes being >10, statistically significantly higher than the average expression of all genes (t-test $p<0.01$). These results support the absence of any major effect of long-term storage.

Epigenomics

Phase I: We did not find any effect of anticoagulant or storage temperature on the yield or quality of isolated DNA or on CpG methylation levels (data not shown). We evaluated the effects of bench-time using buffy coats of 4 subjects. PCA based on M-values shows clear separation between the subjects (Figure 2A) while, in contrast to the corresponding transcriptomics result, no time-dependent trend was evident in PC1-3 (Figure 2A) or other PCs (not shown). The mean coefficient of variation between corresponding probes with $0.01<\beta<0.99$ (thus limited to avoid spurious variability at very low signal intensities) in a 0h vs 8h comparison was 12.2%, not significantly different from that between technical replicates (13.2%). In an ANOVA across the 4 time points, with an additional implementation of a threshold of 20% minimum variation in β, only 3,086 CpG sites (0.6% of the total) showed significant ($p<0.05$) time-dependent variation. STEM analysis of this dataset did not reveal a dominant time-pattern, while overlap between the 4 subjects was minimal (data not shown), strongly suggesting that this variation does not reflect a systematic cellular response.

Phase II: DNA isolated from 42 EPIC-Italy and 38 NSHDS biobank samples was of good quality (OD260/280=1.75-1.85, MW>40,000 kD) and yields were comparable with those obtained with fresh material. We evaluated the suitability of this DNA for microarray-based analysis of CpG methylation by comparing 4 samples from EPIC-Italy and 4 samples from Phase I buffy coats. The fraction of good probes was >99.85% in all cases and only 0.069% of the probes had detection p>0.05 in more than 1 sample and were thus completely excluded. Similar β-value distributions were observed in Phase I and biobank samples (Figure 2B).

Although PCA of 31 samples from each biobank, stored for 13-17 years, shows some scatter (e.g. for samples collected in 1997 in EPIC-Italy and 1996 in NSHDS - 13 and 14 years in storage, respectively), no systematic trend is evident in relation to the storage time (Figure 2C,D). ANOVA indicates that only 50 CpG sites in EPIC-Italy and 1 site in NSHDS samples showed significant variation (Bonferroni-adjusted p<0.05) in methylation levels in relation to storage time.

Metabolomics

Phase I: Of the spectral features detected in all samples analysed for different experimental conditions, 85.9% exhibited a RSD<30% (median RSD=13%) across the QC samples consisting of identical aliquots of a pooled sample interspersed within the batch of regular samples (see Supplemental Material, Methods, Metabolomics). A PCA plot based on these "robust" features indicates a clear separation according to anticoagulant regardless of the donor and other variables (Fig. 3A). For a given anticoagulant, the main sources of variation were the donors and bench-time (Fig. 3B/C, heparin samples only; similar plots were obtained for EDTA and citrate plasma samples), with the 8 and 24 hr timepoints separating away from the earlier timepoints. No general trend was observed in relation to the storage temperature (Fig 3D). The median RSDs of robust peaks reflecting variation by anticoagulant and subject were 11.7% and 18%, respectively, while the effect of bench-time was much smaller and that of storage temperature minimal (Table 3). The numbers of peaks that varied significantly (ANOVA) with both anticoagulant and bench-time were substantially larger than expected by false discovery (71% of peaks at 2% FDR and 6% of peaks at 8% FDR, respectively), confirming the importance of these factors but also that bench-time significantly affected only a relatively small number of metabolites. Similar analysis confirmed that the number of peaks affected by storage temperature (<1%) was below that expected by false discovery.

Phase II: To evaluate the effect of storage time we analysed samples from the same set of subjects as used for the other omics platforms (24 EPIC-Italy and 28 NSHDS plasma samples were available). PCA does not show any systematic effect of storage time (Fig. 3E/F). Overall 72.4% (NSHDS) and 77.2% (EPIC-Italy) of spectral features exhibited an RSD<30% across QC samples. The variation of these robust features across all biobank samples was 2-3fold greater than that associated with storage time (Table 4). ANOVA and false discovery analysis confirmed the absence of a statistically significant association between metabolite peaks and storage time.

Wide-target plasma proteomics

Phase I: Owing to the small number of features measured, only 2 significant principal components were observed. Figure 4A shows that the greatest variation was attributable to the donor, while in addition separation was observed a) by anticoagulant (Figure 4B), with citrate resulting in higher levels of IL2 and IL6 and heparin resulting in higher levels of IL8 (results not shown) and b) by bench-time (Figure 4C), with the 8 and 24hr timepoints deviating most from the earlier ones which were relatively similar. No effect of storage temperature was observed (Figure 4D). The coefficient of variation between different anticoagulants (citrate vs heparin; median 17%; EDTA vs. heparin: median -2.0%) was substantially larger than that between technical replicates (median 2.2%). The latter was similar to the coefficient of variation for the 0 vs 2h bench- time comparison (median -3.0%) and comparable for most analytes to that for 0 vs 4h (median 10.5%; most variation being due to one outlying analyte). However the variation was substantially increased for the 0 vs 8h and even more for the 0 vs 24hr comparison where it reached a median value of 77%.

Phase II: PCA based on the same sets of 31 subjects from each biobank as used with the other platforms does not reveal any systematic effect of storage period (Figure 4E) or collection-to-storage time (data not shown), nor were any associations found with any of the individual analytes. The measured cytokine levels were in the same range as observed in the Phase I (results not shown), suggesting comparability between fresh and biobanked samples.

Discussion

We have evaluated the influence of collection and storage conditions of buffy coat and plasma on sample performance in a series of omics assays, using freshly collected samples as well as samples stored in biobanks for nearly 2 decades. The key findings can be summarised as follows:

Transcriptomics: Transcriptomics-quality RNA can in general be isolated from buffy coat frozen in the absence of RNA preservative, by thawing in the presence of RNAlater, on condition that the buffy coat had been deep-frozen within 8hr of blood collection. No systematic influence of anticoagulant (heparin, EDTA, citrate), storage temperature (-80^{o}C, liquid nitrogen) or time in cold storage on RNA yield or quality (although slightly higher yields were obtained with -80^{o}C and citrate samples) or on the quality of microarray data obtained, was observed. For unknown reasons, a small fraction (<20%) of biobank samples satisfying the above criteria yielded RNA of quality inappropriate for transcriptomics analysis. The majority of samples had RINs between 6 and 8, which, despite indicating a slight degree of RNA degradation, is of more than sufficient quality for transcriptomic analysis (Beekman et al. 2009). Differences in gene expression profiles were mainly observed between different bench-times, followed by anti-coagulants (mainly EDTA vs. heparin) and, to a much lesser extent, storage temperatures. Although it may be possible to compensate for such effects in downstream data analysis by appropriate statistical methods, this observation does underline the importance of recording these variables in biobanks. No systematic effect of time in cold storage on the transcriptomic profiles could be detected, though the latter was studied only in the rather limited range of 13-17 years.

Epigenomics (CpG methylation): DNA suitable for microarray-based analysis of CpG methylation levels can be obtained from biobank buffy coats frozen within 8 hours of blood collection. No systematic influence of anticoagulant, storage temperature or length of cold storage in the biobank (over the period examined) on DNA yield or quality or methylation profiles was observed. Bench-time appears to affect methylation levels at a very small fraction (0.6%) of the CpG sites in a non-systematic way and its overall impact on the information content of the resulting data would be very limited.

Plasma UPLC-ToFMS metabolomics: Unlike DNA or RNA, no universal indicator of "quality" can be defined for the metabolome, where each molecule detected exhibits a different stability profile. Hence the impact of collection and storage conditions on the metabolomic profile is difficult to define comprehensively. Using multivariate analysis we could detect no significant influence of storage temperature or length of cold storage in the biobank within the storage period examined. Although for all anticoagulants used good quality data were obtained, the metabolomic profiles were strongly influenced by the anticoagulant employed. From a technical perspective, heparin is preferable over citrate or EDTA which can reduce column lifetime and increase ion suppression. Bench-time affected only a minor fraction of the profile but with substantial changes occurring beyond 4hr. Other studies using NMR (Barton et al 2008) and GC-MS (Dunn et al 2008) have shown that plasma samples are stable at 4^{o}C for up to 24hr. While we consider our findings to be broadly consistent with other reports (Dunn et al 2011; Zelena et al 2009), some have reported more robust features using UPLC-QTOF-MS analysis of serum (e.g. Dunn et al. 2011 report 83.9%+/-3.1 of peaks with an RSD<20% across QCs). Key differences between these studies and ours include the use of serum versus plasma, the precise detector used and, importantly, the application of LOESS regression to correct for technical peak intensity variation.

Wide-target proteomics: Plasma in long-term storage can be successfully subjected to proteomic analysis, provided that it was isolated and frozen within 4 hours of blood collection. No influence of storage temperature or length of long-term cold storage in biobanks on the corresponding profiles was observed over the period examined. However, a major influence of the anticoagulant was observed, in line with an earlier report (Saberi Hoshnije et al. 2010) in which strong correlations were observed between heparin and citrate plasma although small differences in analyte levels were observed for most analyses (11 cytokines, 4 chemokines, and 2 adhesion molecules).

Conclusions

The overall conclusion that can be drawn from these findings is that it appears likely that a large fraction of the human blood-derived samples currently in long-term storage in biobanks is amenable to analysis using high-throughput omics technologies, even if no precautions specifically related to the eventual use of these technologies were taken at the time of collection. Important criteria which should be considered in selecting samples (including freshly collected samples) for such analyses are a) time between blood collection and fractionation should not exceed 8hr (4hr for proteomics), and b) samples whose data are to be compared or pooled should not contain different anticoagulants. Although an influence on omic profiles of additional variables, especially the length of time in cold storage, cannot be precluded owing to the relatively limited span of years in storage evaluated, adherence to these criteria minimises the impact of sample history and facilitates the generation of reliable data. Within these limitations, interindividual differences were found to be by far the largest source of variation in omic profiles of biosamples. As previously noted, these profiles (e.g. in the blood transcriptome) can reflect the corresponding profiles in other tissues and the effects thereupon of environmental factors (Liew et al. 2006). These findings open the way to the application of these powerful technologies to biosamples collected over previous decades in the context of population-based or disease-oriented cohorts. In combination with other available information from many such cohorts (e.g. environmental exposure, dietary or life-style habits, disease status, or related biomarkers), such application is likely to provide strong support to research on the environmental causes of disease.

References

Barton RH, Nicholson JK, Elliott P, Holmes E. 2008. High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study. Int J Epidemiol 37 Suppl 1:i31-40.

Beekman JM, Reischl J, Henderson D, Bauer D, Ternes R, Peña C et al. 2009. Recovery of microarray-quality RNA from frozen EDTA blood samples. J Pharmacol Toxicol Methods 59:44-49.

Bingham S, Riboli E. 2004. Diet and cancer - the European Prospective Investigation into Cancer and Nutrition. Nat Rev Cancer 4:206-215.

Chatziioannou A, Moulos P, Kolisis FN. 2009. Gene ARMADA: an integrated multi-analysis platform for microarray data implemented in MATLAB. BMC Bioinformatics 10: 354; doi: 10.1186/1471-2105-10-354.

Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11:587; doi:10.1186/1471-2105-11-587.

Dunn WB, Broadhurst D, Ellis DI, Brown M, Halsall A, O'Hagan S et al. 2008. A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. Int J Epidemiol. 37 Suppl 1:i23-30.

Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N et al. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 6:1060-1083.

Ellinger-Ziegelbauer H, Aubrecht J, Kleinjans JC, Ahr HJ. 2009. Application of toxicogenomics to study mechanisms of genotoxicity and carcinogenicity. Toxicol Lett. 186:36-44.

EnviroGenomarkers Project: Genomics Biomarkers of Environmental Health. n.d. Available from: http://www.envirogenomarkers.net.

Ernst J, Bar-Joseph Z. 2006. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics 7:191; doi:10.1186/1471-2105-7-191.

Hallmans G, Agren A, Johansson G, Johansson A, Stegmayr B, Jansson JH et al. 2003. Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. Scand J Public Health Suppl. 61:18-24.

Karlovich C, Duchateau-Nguyen G, Johnson A, McLoughlin P, Navarro M, Fleurbaey C et al. 2009. A longitudinal study of gene expression in healthy individuals. BMC Med Genomics 2:33; doi: 10.1186/1755-8794-2-33.

Liew C-C, Ma J, Tang H-C, Zheng R, Dempsey AA. 2006. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J Lab Clin Med 147:126–132.

Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK et al. 2004. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Prospect 112:1691–1696.

Mancuso FM, Montfort M, Carreras A, Alibes A, Roma G. 2011. HumMeth27QCReport: an R package for quality control and primary analysis of Illumina Infinium methylation data. BMC Res Notes. 4 4: 546; doi:10.1186/1756-0500-4-546.

McHale CM, Zhang L, Hubbard AE, Smith MT. 2010. Toxicogenomic profiling of chemically exposed humans in risk assessment. Mutat Res. 705:172-83

Pereira H, J-FM, Joly C, Sebedio J-L, Pujos-Guillot E. 2010. Development and validation of a UPLC/MS method for a nutritional metabolomic study of human plasma. Metabolomics 6:207-218.

Rappaport SM, Smith MT. 2010. Environment and disease risks. Science 330:460–461.

Saberi Hosnijeh F, Krop EJ, Scoccianti C, Krogh V, Palli D, Panico S et al. 2010. Plasma cytokines and future risk of non-Hodgkin lymphoma (NHL): a case-control study nested in the Italian European Prospective Investigation into Cancer and Nutrition. Cancer Epidemiol Biomarkers Prev. 19:1577-1584.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M et al. 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 7:3; doi:10.1186/1471-2199-7-3

Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P et al. 2009. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. Anal Chem. 81:1357-1364.

Table 1: RNA integrity numbers (RINs) and RNA yields (µg) of fresh samples from 4 subjects (mean ± sd) according to anticoagulant, storage temperature and bench time (0—24 hr).

Anticoagulant, storage temp.	**0hr**		**2hr**		**4hr**		**8hr**		**24hr**	
	RIN	**yield**	**RIN**	**yield**	**RIN**	**yield**	**RIN**	**yield**	**RIN**	**yield**
Citrate										
-80°C	7.15±0.14	8.27±2.23	7.25±0.42	8.90±0.53	7.43±0.53	8.75±1.59	7.70±0.21	8.29±1.64	7.38±0.24	12.19±4.93
liq. N_2	7.10±0.49	5.96±1.43	7.68±0.11	5.74±0.80	7.65±0.14	4.77±0.56	7.33±0.11	6.56±2.77	7.08±0.04	6.62±3.55
EDTA										
-80°C	6.43±0.88	5.14±2.20	6.53±0.88	10.46±7.33	7.30±0.21	5.21±2.93	7.50±0.00	6.98±7.06	7.60±0.35	5.50±3.17
liq. N_2	6.75±0.64	3.29±1.26	6.95±0.92	4.09±2.61	7.55±0.07	4.24±3.05	7.18±0.04	3.95±1.80	7.20±0.14	5.22±1.16
Heparin										
-80°C	6.95±0.21	4.81±0.43	5.28±2.93	6.48±5.23	7.50±0.00	7.06±4.14	6.78±1.31	6.70±3.15	7.93±0.18	6.51±1.60
liq. N_2	6.68±0.95	3.20±0.97	6.88±0.95	4.24±2.17	7.53±0.18	3.93±2.62	7.45±0.35	4.09±0.15	7.50±0.00	3.73±0.99

Yields were obtained from 0.4-0.5 ml of buffy coat (corresponding to ~2 ml blood). -

Table 2: Average RNA integrity numbers (RINs) and RNA yield (µg) from biobank samples

Cohort	n	% RIN > 6	RIN (average ± sd)	RNA yield (mean ± sd)
EPIC- Italy	64	95	7.1 ± 0.7	3.9 ± 1.7
NSHDS	50	92	7.4 ± 0.9	12.2 ± 7.5

The EPIC-Italy sample set included 6 samples stored at -80°C with a RIN of 6.8 ± 0.5 and RNA yield of 5.1 ± 1.2. The remaining samples were stored in liquid N_2. The NSHDS sample set included 9 samples with EDTA as anti-coagulant with a RIN of 6.7 ± 0.8 and RNA yield of 13.9 ± 6.8 µg. The remaining samples used heparin. EPIC-Italy and NSHDS yields were obtained from half a cryostraw or half an eppendorf of buffy coat, corresponding to approximately 0.25 and 0.7-1.0 ml buffy coat (corresponding to ~3 and ~9 ml blood), respectively.

Table 3. Relative standard deviation (RSD) of metabolomics peaks across experimental conditions in Phase I

Conditions	**RSD of samples within different percentiles**		
	10th %	Median	90th %
QC samples	7.0%	13.0%	37.0%
Subjects[a]	13.0%	18.0%	49.4%
Anticoagulants [a,b]	4.5%	11.7%	46.6%
Storage temperatures [a,b]	0.4%	1.8%	5.0%
Bench-times [a,b]			
0 vs. 2hr	0.6%	2.7%	7.4%
0 vs. 4hr	1.0%	2.8%	7.1%
0 vs. 8hr	1.7%	3.5%	9.3%
0 vs. 24hr	2.2%	4.8%	15.4%

RSDs were calculated by comparing samples differing in the condition indicated while keeping all other conditions constant; "QC samples" refers to comparison across identical quality control samples; "bench-times" refers to comparison between samples with bench-time 0hr and the time indicated;

[a]using only selected peaks with RSD<30% across QCs

[b]data were normalised to mean value of donor

Table 4. Relative standard deviation (RSD) of metabolomics peaks across subjects and storage times for Phase II samples

	RSD of samples within different percentiles		
	10th %	**Median**	**90th %**
EPIC-Italy			
all subjects	12.8%	26.2%	55.8%
storage time	3.9%	10.8%	24.3%
NSHDS			
all subjects	13.8%	27.8%	54.6%
storage time	3.9%	8.8%	23.1%

For each cohort, RSDs were calculated by comparing a) samples of all subjects and b) the means of samples with the same storage time

Figure Legends

Figure 1. Transcriptomics. A-D, Phase I – data from 4 different subjects (samples from same subjects are indicated with same symbols); A: PCA plot on samples with different bench-times (indicated in hr by the symbol labels) (EDTA, -80 °C; proportion of variance explained: PC1 31%, PC2 24%, PC3 14%). B: PCA on same samples but using PC4 instead of PC3 (proportion of variance explained: PC4 10%); the line indicates the bench-time-related trend. C: PCA on samples with different anticoagulants (bench-time 2.5 hr, -80°C; proportion of variance explained: PC1 57%, PC2 13%, PC3 10%). D: PCA on samples with different storage temperatures (EDTA, bench-time 0 hr; proportion of variance explained: PC1 39%, PC2 21%, PC3 19%). E-H, Phase II; E,F: Comparison of microarray data from 4 fresh and 4 biobank samples. Average intensity level after LOESS and A-quantile normalization (E), numbers of good array probes (F); G,H: PCA plots on storage time in biobank; the legend indicates the number of years in storage (G: EPIC-Italy, proportion of variance explained: PC1 32%, PC2 13%, PC3 8%; H: NSHDS, proportion of variance explained: PC1 40%, PC2 10%, PC3 8%).

Figure 2. Epigenomics. A: Phase 1, PCA plot on bench-times from 4 subjects (samples from same subjects are indicated by same symbols, labels indicate bench-time in hrs) (EDTA, -80°C; proportion of variance explained: PC1 37%, PC2 18%, PC3 16%). B: Distribution of beta values from 4 fresh and 4 EPIC-Italy samples after LOESS and A-quantile normalization. C,D: Phase 2, PCA plots on storage time (number of years) in biobank (C: EPIC Italy; proportion of variance explained: PC1 20%, PC2 10%, PC3 7%; D: NSHDS, PC1 19%, PC2 13%, PC3 6%).

Figure 3. Metabolomics. A-D, Phase I; A: PCA plot on anticoagulants from 4 subjects; because all samples were subjected to full metabolomics analyses, the points shown for each anticoagulant include different subjects, bench-times and storage temperatures (proportion of variance explained: PC1 44%, PC2 20%, PC3 10%). B: PCA on bench-time for 3 subjects (different symbols denote different subjects and include 2 different storage temperatures per subject; the labels denote bench-times in hr (only 0h bench-time for 1 subject) (heparin; proportion of variance explained: PC1 56%, PC2 20% PC3 12%). C: PCA for same samples as D but PC4 instead of PC3 (proportion of variance explained: PC4 6%); the line indicates the bench-time-related trend. D: PCA on storage temperature from 4 subjects (different symbols denote different subjects) (heparin, bench-time 0 hr; proportion of variance explained: PC1 51%, PC2 31%, PC3 14%). E-F, Phase II; E,F: PCA on storage time (years) in biobank (E: EPIC-Italy; proportion of variance explained: PC1 35%, PC2 10%, PC3 8%; F: NSHDS; proportion of variance explained: PC1 36%, PC2 21%, PC3 7%).

Figure 4. Proteomics. A-D: Phase 1. PCA plot (proportion of variance explained: PC1 56.4%, PC2 25.0%) labelled for donor (A), storage temperature (B), bench-times (C), anticoagulants (D). Because all samples were subjected to proteomic analyses, the points shown for each variable indicated include variation of the remaining variables. E: PCA on storage time in the two biobanks.

Figure 1

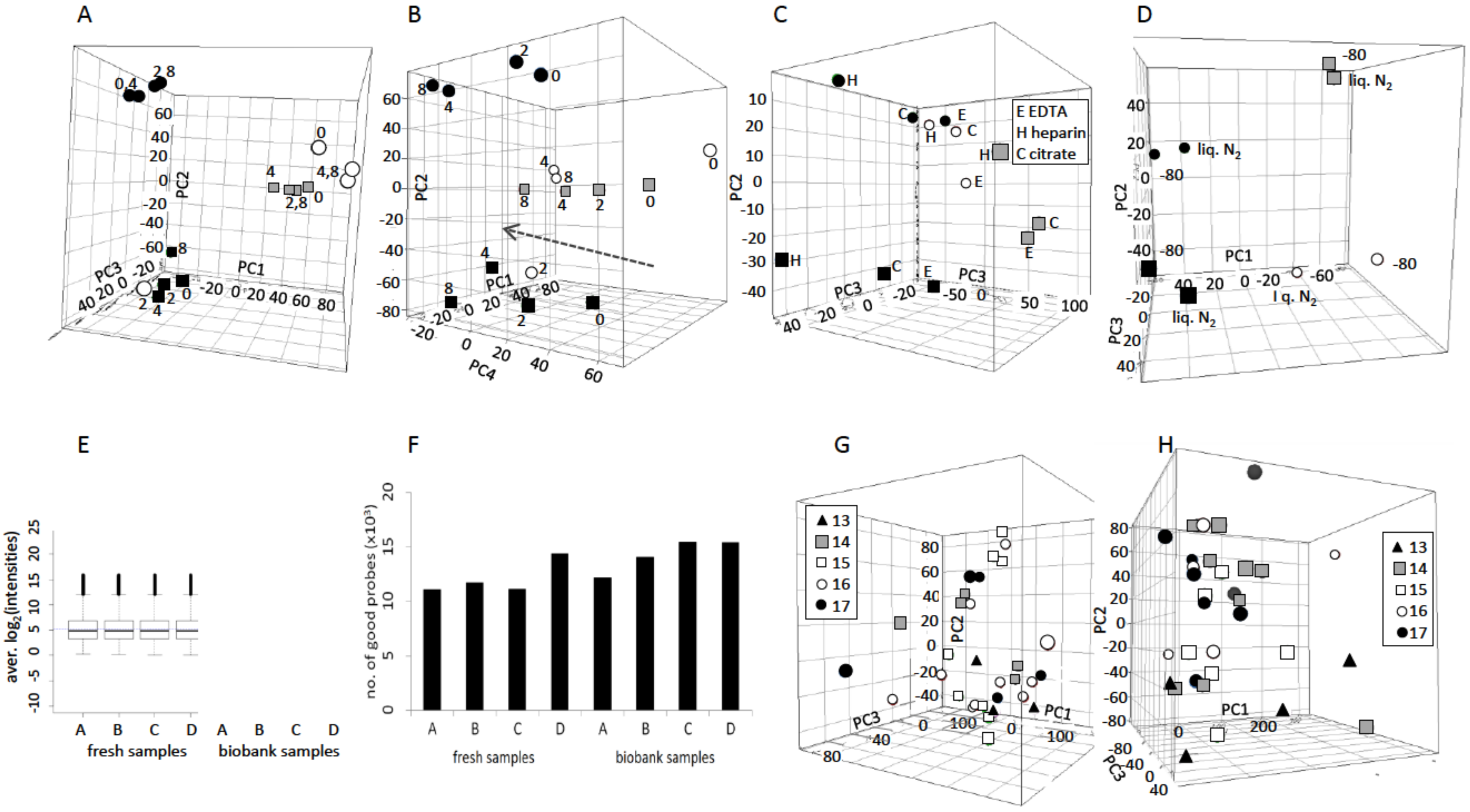

A
PC1
PC2
PC3
0,4
2 8
0
4,8
4
2,8
0
8
2
2
0
4
B
PC1
PC2
PC4
8
4
2
0
4
8
8
4
2
0
0
4
2
8
2
0
C
E EDTA
H heparin
C citrate
PC2
PC3
H
C
H
E
C
H
E
E
C
H
C
E
D
PC1
PC2
PC3
-80
liq. N2
liq. N2
-80
-80
l iq. N2
liq. N2
E
aver. log2(intensities)
A
B
C
D
A
B
C
D
fresh samples
biobank samples
F
no. of good probes (x10³)
A
B
C
D
A
B
C
D
fresh samples
biobank samples
G
▲ 13
■ 14
□ 15
○ 16
● 17
PC1
PC2
PC3
H
▲ 13
■ 14
□ 15
○ 16
● 17
PC1
PC2
PC3

Figure 2

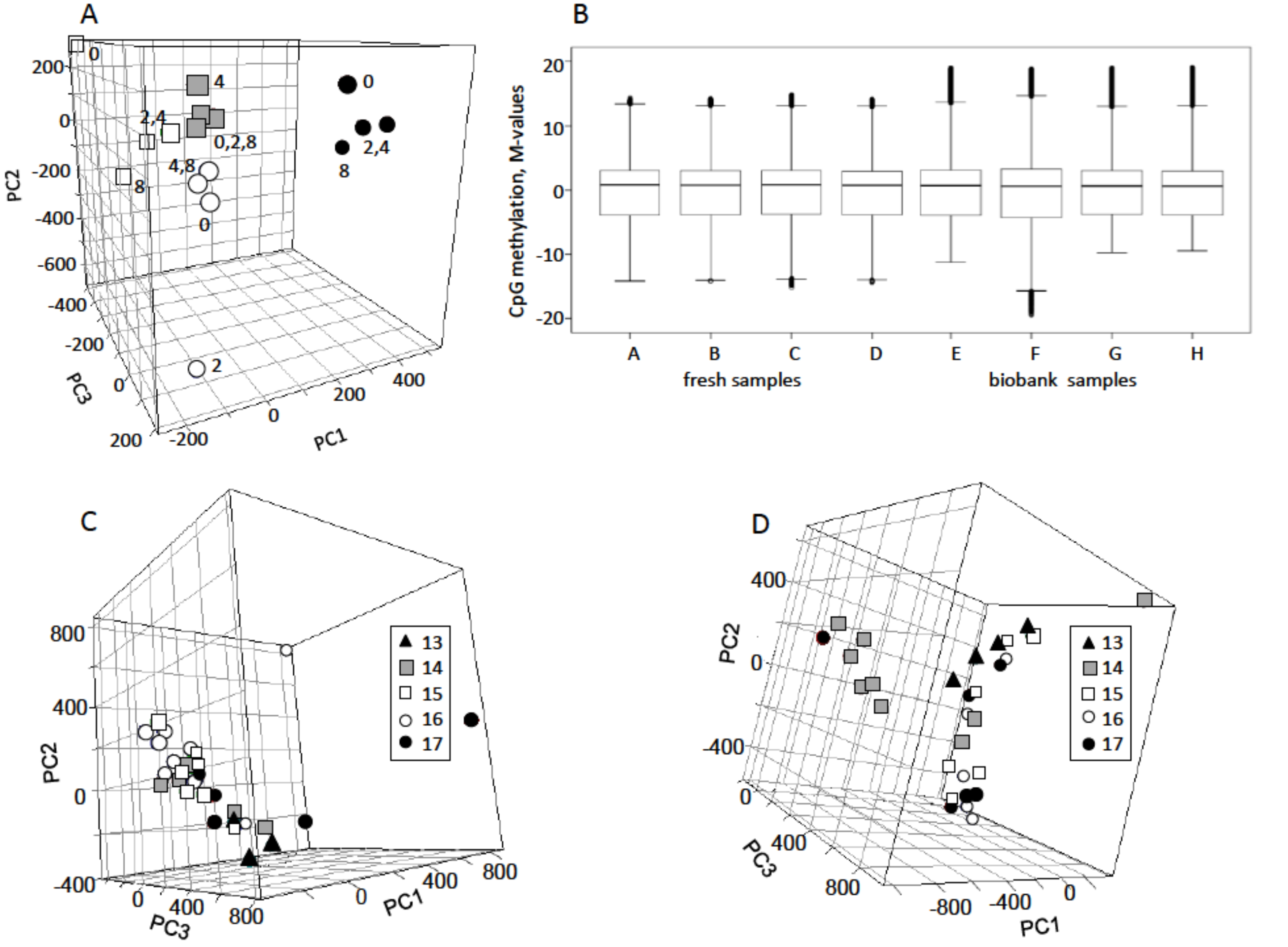

A
0
4
0
2,4
0,2,8
2,4
4,8
8
8
0
2
200
0
-200
-400
-600
PC2
-400
-200
0
200
PC3
-200
0
200
400
PC1
B
20
10
0
-10
-20
CpG methylation, M-values
A
B
C
D
E
F
G
H
fresh samples
biobank samples
C
800
400
0
-400
PC2
0
400
800
PC3
0
400
800
PC1
▲ 13
■ 14
□ 15
○ 16
● 17
D
400
0
-400
PC2
0
400
800
PC3
-800
-400
0
PC1
▲ 13
■ 14
□ 15
○ 16
● 17

Figure 3

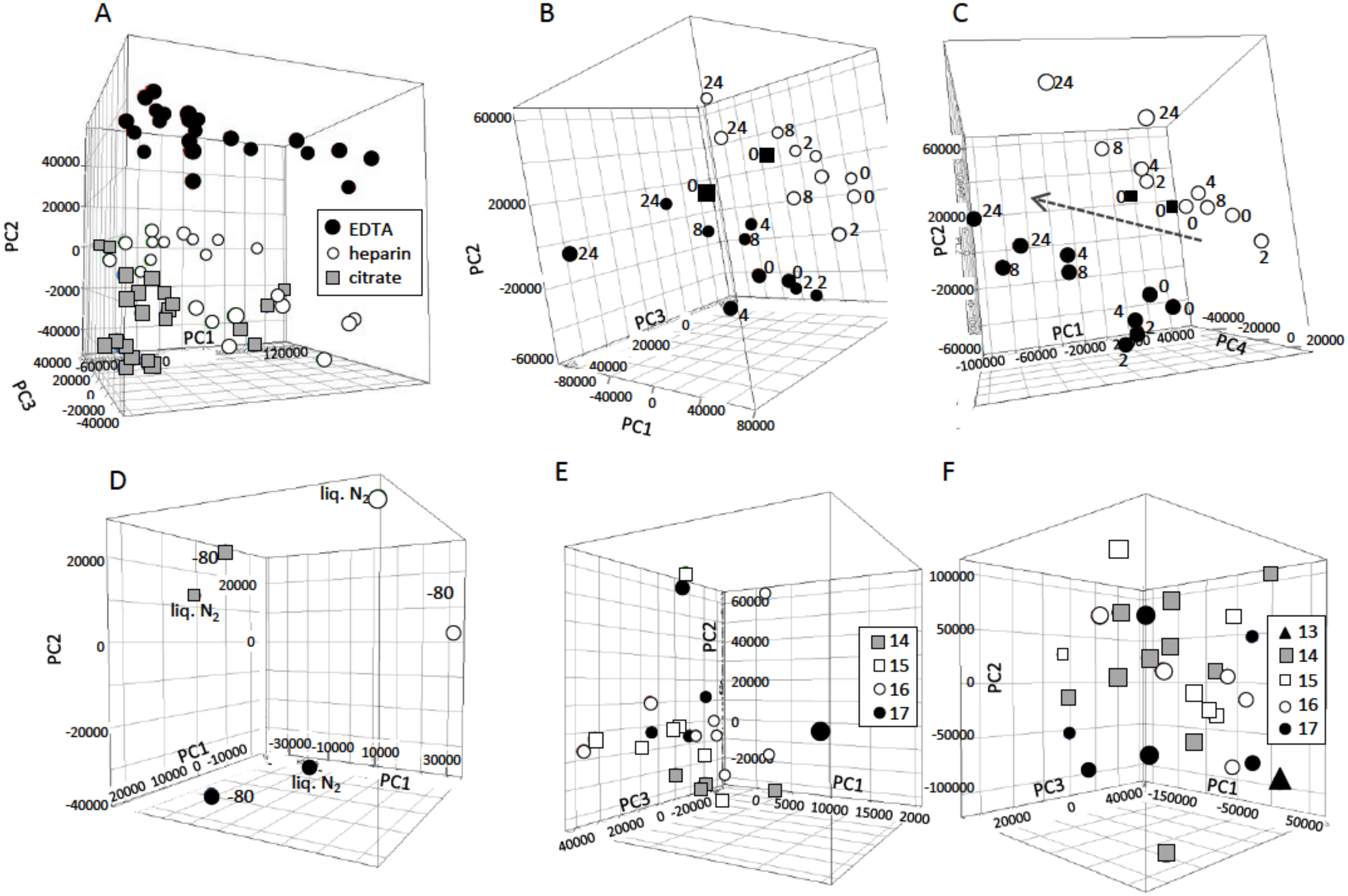
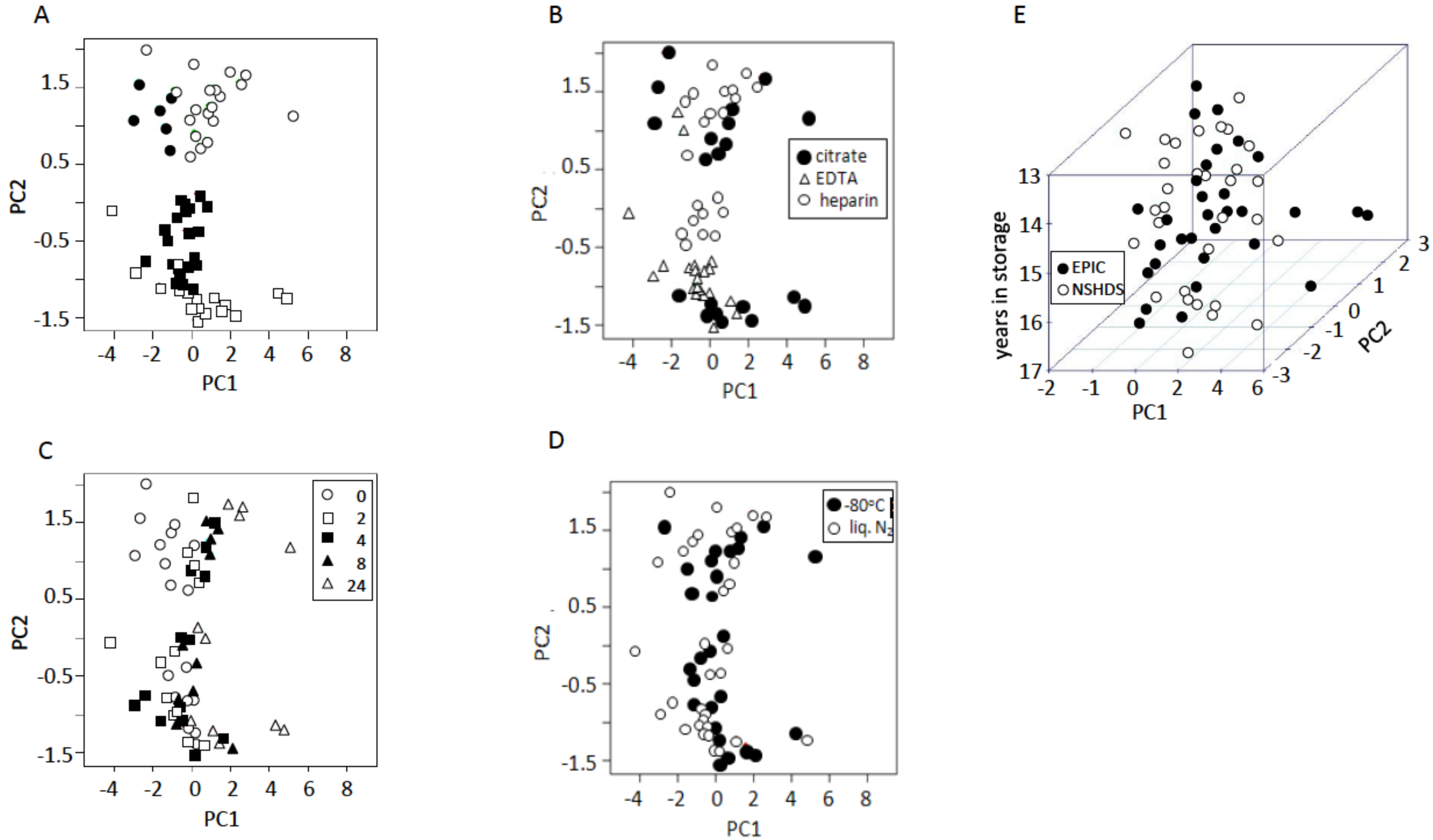

A
EDTA
heparin
citrate
PC1
PC2
PC3
40000
20000
0
-20000
-40000
-60000
120000
B
24
8
2
0
4
PC1
PC2
PC3
60000
20000
-20000
-60000
-80000
-40000
40000
80000
C
24
8
4
2
0
PC1
PC2
PC4
60000
20000
-20000
-60000
-100000
-40000
0
20000
D
liq. N2
-80
PC1
PC2
20000
0
-20000
-40000
-30000
-10000
10000
30000
E
14
15
16
17
PC1
PC2
PC3
60000
40000
20000
0
5000
10000
15000
F
13
14
15
16
17
PC1
PC2
PC3
100000
50000
0
-50000
-100000
-150000
50000

Figure 4


A
PC2
1.5
0.5
-0.5
-1.5
-4 -2 0 2 4 6 8
PC1
B
PC2
1.5
0.5
-0.5
-1.5
-4 -2 0 2 4 6 8
PC1
● citrate
△ EDTA
○ heparin
E
years in storage
13
14
15
16
17
-2 -1 0 2 4 6
PC1
3 2 1 0 -1 -2 -3
PC2
● EPIC
○ NSHDS
C
PC2
1.5
0.5
-0.5
-1.5
-4 -2 0 2 4 6 8
PC1
○ 0
□ 2
■ 4
▲ 8
△ 24
D
PC2
1.5
0.5
-0.5
-1.5
-4 -2 0 2 4 6 8
PC1
● -80°C
○ liq. N2